Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1

Supplementary Material

M. Figliuzzi,[1,2,3] H. Jacquier,[4,5] A. Schug,[6] O. Tenaillon,[4] M. Weigt*[,2,3]

[1]*Sorbonne Universités, UPMC, Institut de Calcul et de la Simulation, Paris, France*
[2]*Sorbonne Universités, UPMC, UMR 7238, Computational and Quantitative Biology, Paris, France*
[3]*CNRS, UMR 7238, Computational and Quantitative Biology, Paris, France*
[4]*INSERM, IAME, UMR 1137, Paris, France*
Université Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, Paris, France
[5]*Service de Bactériologie-Virologie, Groupe Hospitalier Lariboisiére-Fernand Widal, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France*
[6]*Steinbuch Centre for Computing, Karlsruhe Institute for Technologie, Eggenstein-Leopoldshafen, Germany*
∗***Corresponding author*** *mail: martin.weigt@upmc.fr*

Supplementary Tables S1-S3: Lists of strongly mispredicted mutations

Mutation	Experimental MIC	DCA	IND
F72Y	25	500	500
V216D	12.5	500	500
G251R	12.5	500	500
G251W	12.5	500	1000
L51F	500	12.5	12.5
A248T	1000	12.5	12.5

TABLE S1: **Mutations mispredicted by both IND and DCA modeling**. Underlined mutations fall into highly gapped position of the MSA. Positions are indicated using standard Ambler numbering.

Mutation	Experimental MIC	DCA	IND
F66L	25	500	250
P67L	25	500	250
P257L	500	25	50

TABLE S2: **Mutations mispredicted by DCA and (partially) corrected by IND modeling**. Underlined mutation falls into highly gapped position of the MSA. Positions are indicated using standard Ambler numbering.

Mutation	Experimental MIC	DCA	IND
D179N	25	250	500
A237D	12.5	250	500
I246N	12.5	50	500
G251E	12.5	250	1000
G54A	500	100	25
E63V	500	100	12.5
T114M	500	250	12.5
N154Y	500	500	12.5
A185V	500	50	12.5
T188P	500	250	12.5
D209V	500	250	25
D254Y	500	50	12.5

TABLE S3: **Mutations mispredicted by IND and (partially) corrected by DCA modeling**. Underlined mutations fall into highly gapped position of the MSA. Positions are indicated using standard Ambler numbering.

Supplementary Figure 1

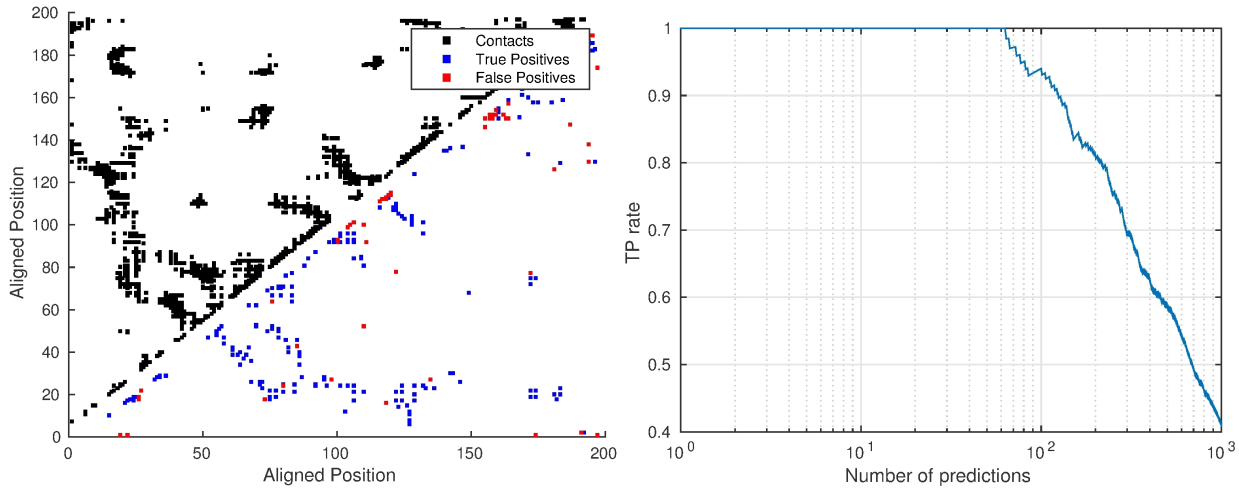

FIG. S1: **Contact prediction result for TEM-1:** Contact map predictions using standard DCA analysis. **Left panel** – Black symbols in the upper left side represent native contacts with a cutoff of 8Å, colored symbols in the bottom right side refer to the first 200 not-trivial predictions (obtained computing the average-product correction of the Frobenious Norm of the epistatic couplings ϕ_{ij} for all i, j with $|i - j| > 4$, as detailed in [1]). Among these, true-positive predictions are in blue, false-positive ones in red. **Right Panel** – True Positive rate of contact prediction (on the set of pairs i, j of residues further than 4 aminoacids on the primary sequence, i.e. $|i - j| > 4$) as a function of the number of predicted contacts. The first 63 predictions are without errors.

Supplementary Figure 2

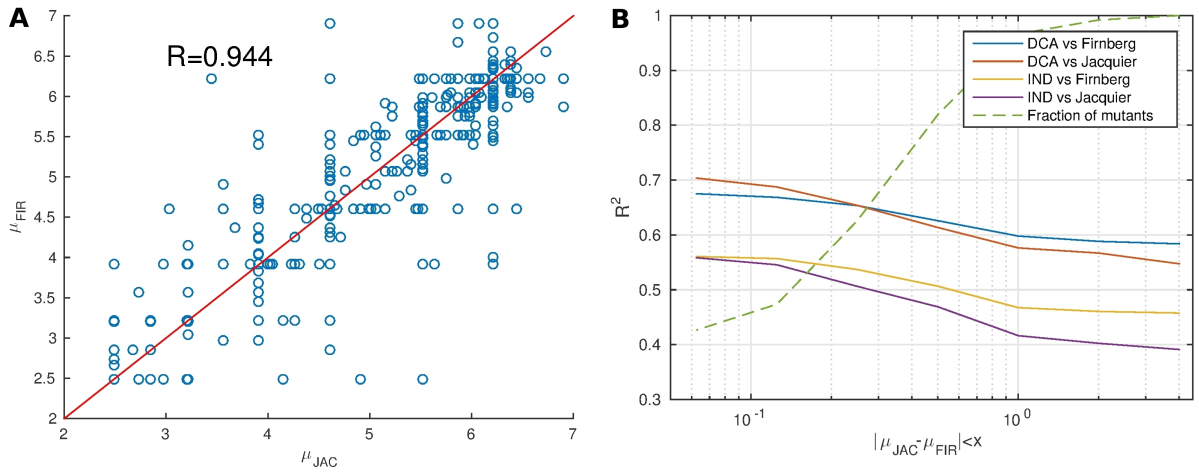

FIG. S2: **Comparison of experimental measures. Panel A** – Scatter plot of MIC measured in [2] (μ_{FIR}) vs. MIC measured in [3] (μ_{JAC}). Since the two measures are in slightly non-linear dependence, mainly due to a difference in their dynamical range, we have previously mapped [2] into [3] as done with computational predictors, to better compare the outcomes of two experiments and to spot those mutations which are significantly differently characterized by the two experiments. After the mapping, the two measures are highly correlated ($R = 0.944$). **Panel B** – Performance excluding mutation displaying experimental discrepancies: R^2 between experimental MIC and predicted one excluding those mutations where the difference in measured impact is greater than a given threshold: $|\mu_{JAC} - \mu_{FIR}| > x$.

Supplementary Figure 3

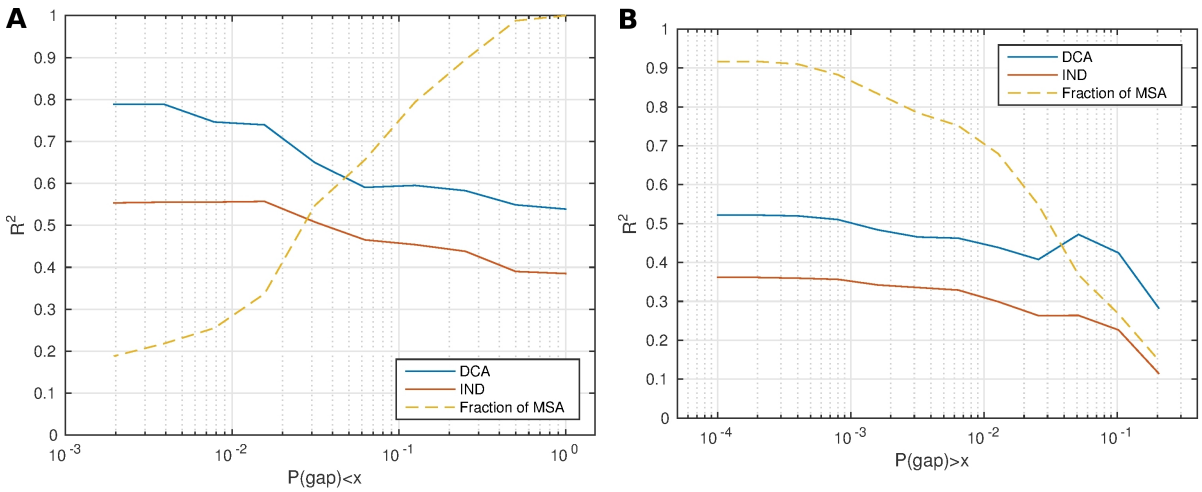

FIG. S3: **Performance of IND model and DCA excluding most gapped and least gapped positions:** Pearson correlation between experimental MIC and predicted ones excluding those mutations falling in positions where the gap frequency in the MSA $P_i(gap)$ is greater then a given threshold x (Panel A) and those in which $P_i(gap)$ is smaller than x (Panel B).

Supplementary Figure 4

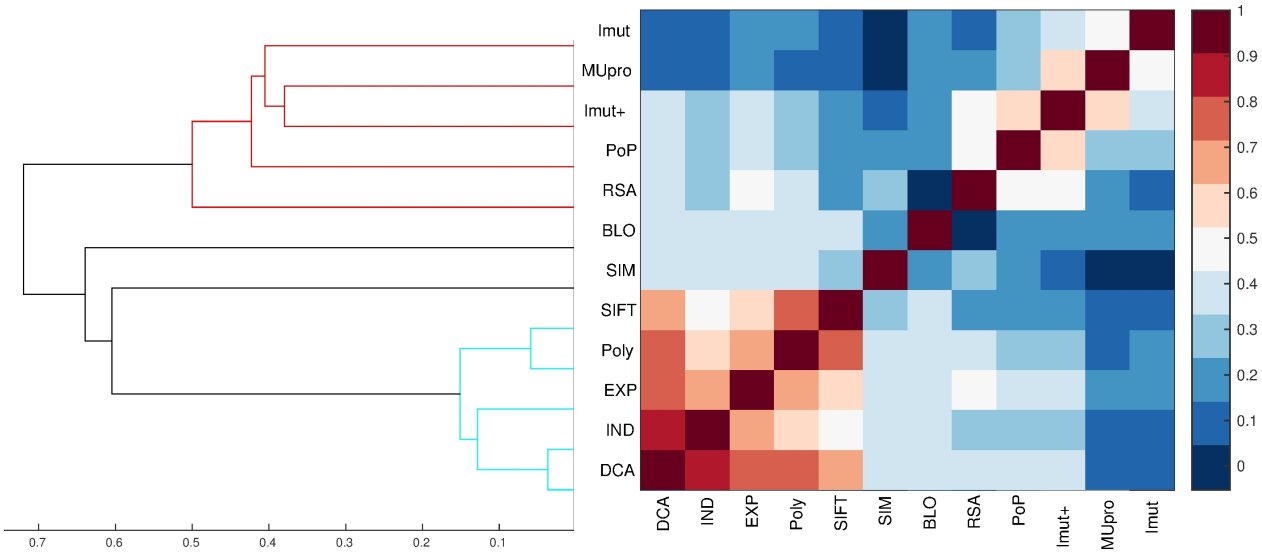

FIG. S4: **Cross-correlations between methods for the beta-lactamase TEM-1:** Pearson correlations between the following quantities: Experimental Fitness (EXP), DCA (DCA), Independent-Site Model (IND), Polyphen-2 (Poly), SIFT (SIFT), Blosum Substitution Matrix (BLO), Molecular Simulations (SIM), Relative Solvent Accessibility (RSA) , PoPMuSiC (PoP), I-Mutant2.0 (Imut), MUpro (MUpro) and I-Mutant2.0(sequence+structure) (Imut+). On the left the hierarchical clustering of the correlation matrix (taking $d = 1 - R$ as metrics). The structure of cross correlation matrix between different predictions reveals cluster of methods sharing the same source of information strongly correlated: one cluster is formed by bioinformatics prediction programs, Blosum and RSA are completely orthogonal, structure-based and evolutionary based approaches are modestly correlated.

Supplementary Figure 5

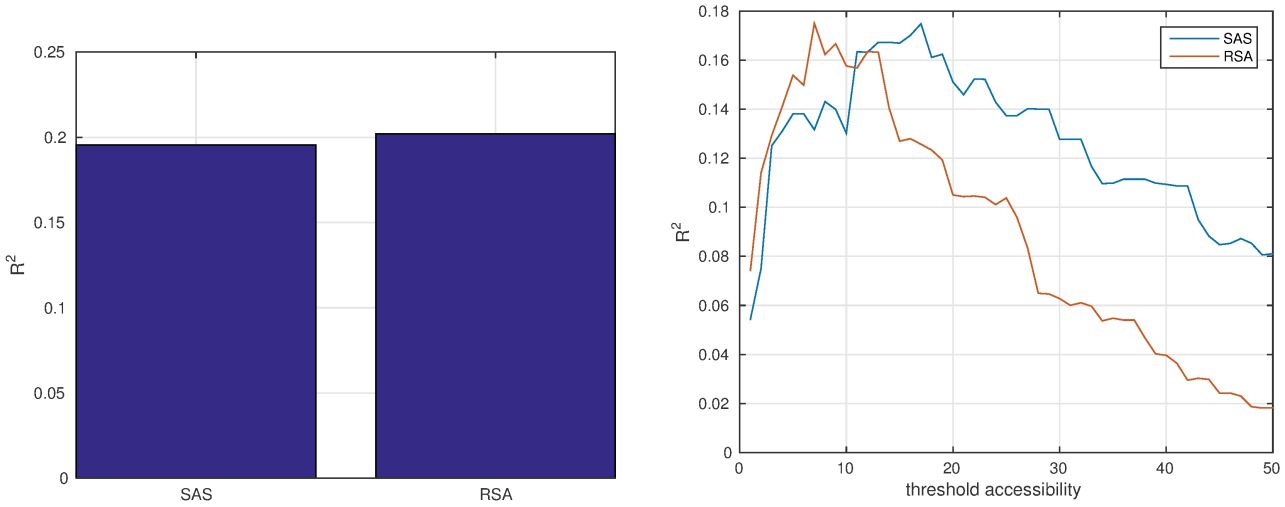

FIG. S5: **Accessibility-based predictions for the beta-lactamase TEM-1. Left panel** – R^2 between experimental fitness Solvent Accessible Surface (SAS) or Relative Surface Accessibility (RSA). **Right Panel** – R^2 between experimental fitness and a binary classifier specifying whether a residue is more exposed than a given threshold, as a fuction of the threshold (measured in Å^2 in the case SAS, and in Percentage of Accessibility in the case of RSA)

Supplementary Figure 6

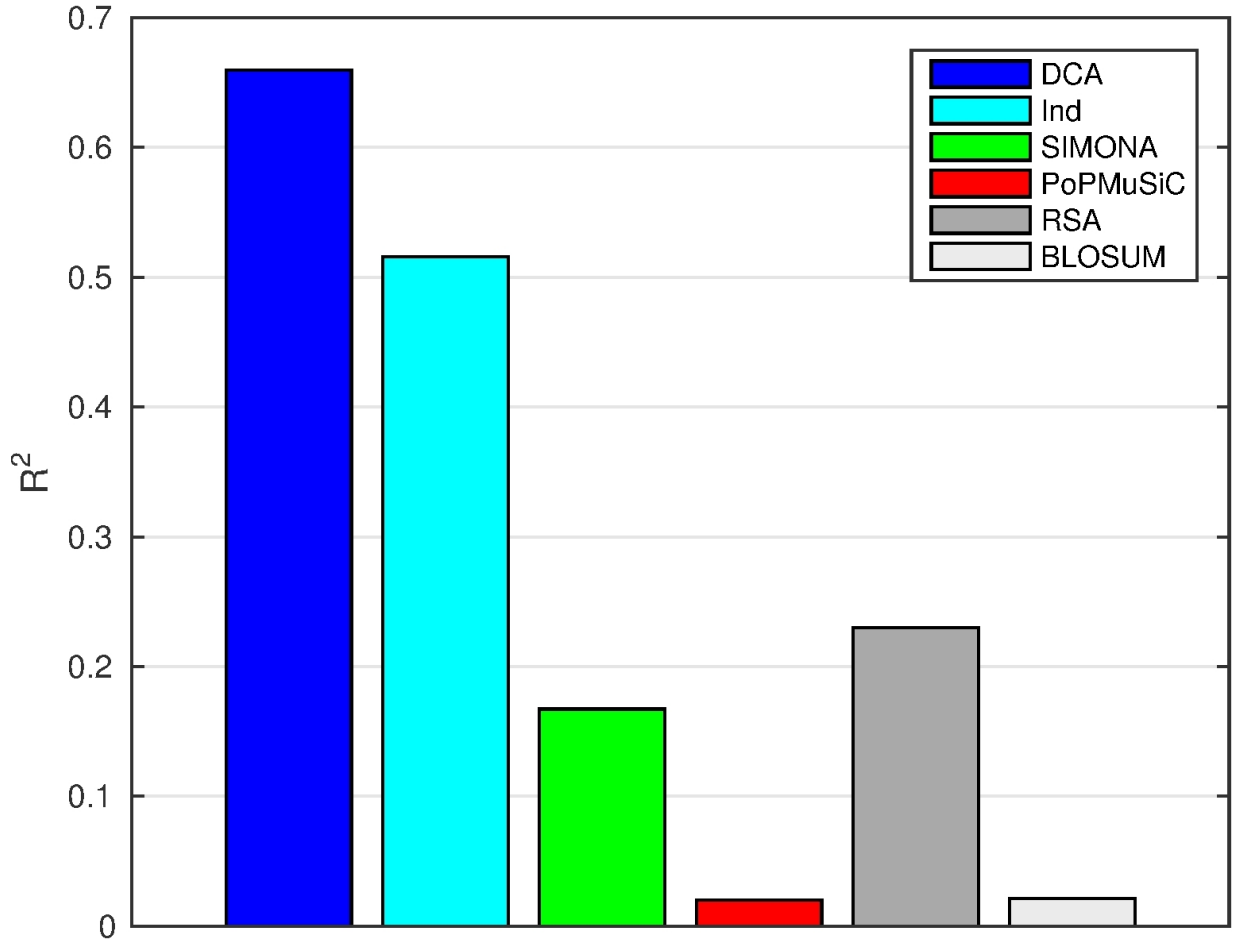

FIG. S6: **Prediction of mutational effects in the active site of the beta-lactamase TEM-1:** R^2 computed on a subset of 111 mutations affecting the extended active site. Our definition of extended active site includes the tetrad Ser70-XX-Lys73, which contains the main catalytic residue (Ser70); the so-called "SDN loop", formed by Ser130, Asp131 and Asn132 residues, since it has been suggested that Ser130 is involved in proton transfer from Ser70 to the β-lactam ring during acylation; the highly conserved Ω loop (residues 161-179), essential for the positioning of the water molecule; the Lys234-Ser235-Gly236 sequence: crystallographic data indicate that Ser130 and Lys234 are connected by a hydrogen bond, which would serve as a connection between the two domains of the protein and help to stabilize the active site. Finally, Asn170, Lys234, Ser235, Ala237 and Arg244 residues Glu104 would also be involved in the stabilization of the acyl enzyme intermediate.

Supplementary Figure 7

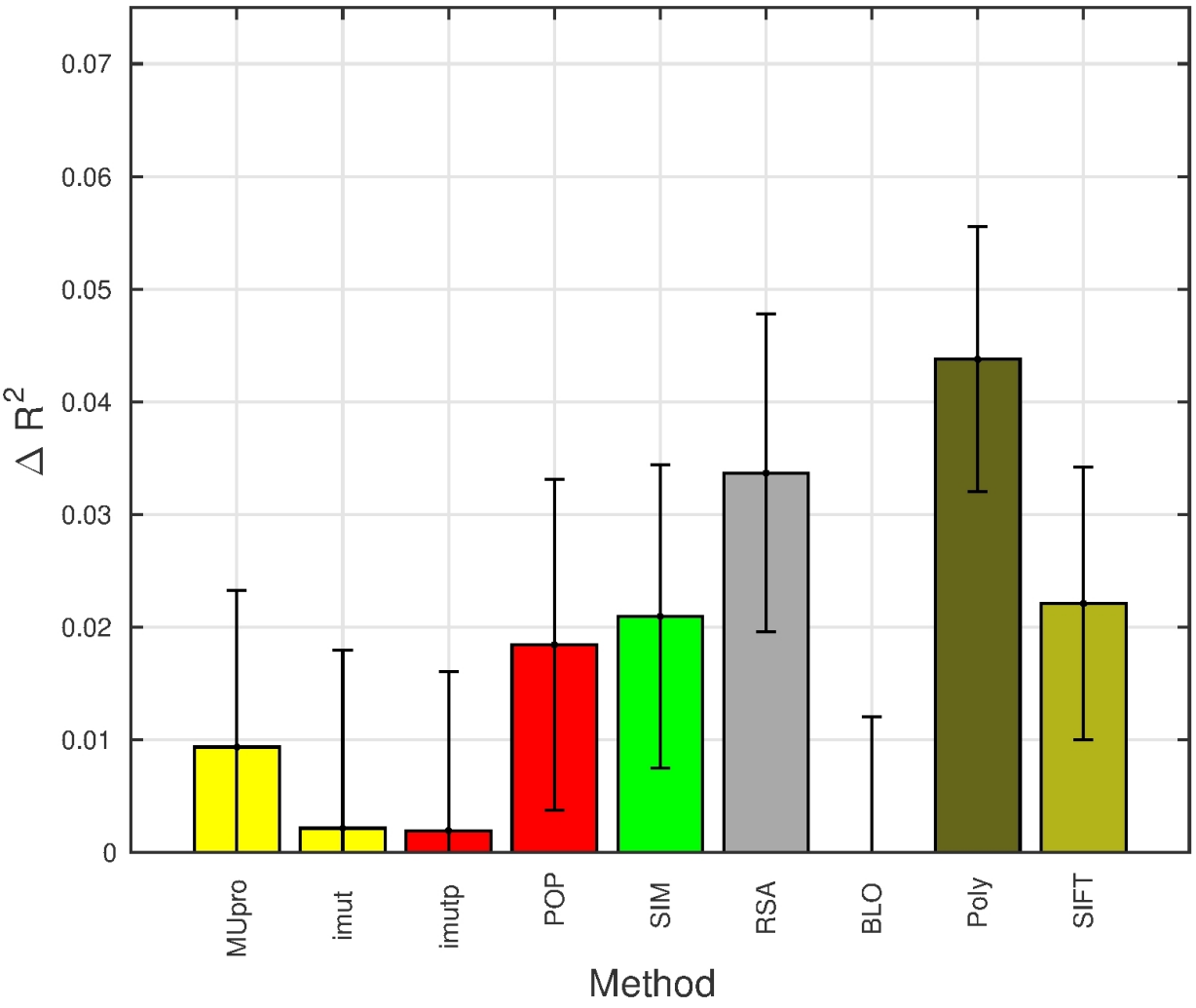

FIG. S7: **Combination of DCA and other predictors for the beta-lactamase TEM-1:** Increase in R^2 given by linear combination of DCA with any of the other methods: MUpro (MUpro), I-Mutant2.0 (Imut), PoPMuSiC (PoP), I-Mutant2.0(sequence+structure) (Imut+), Molecular Simulations (SIM), and Relative Solvent Accessibility (RSA), Blosum Substitution Matrix (BLO), Polyphen-2 (Poly) and SIFT (SIFT). To build and assess the performance of the linear models, we use threefold cross-validation. Error bars indicate average errors in the cross-validation.

Supplementary Figure 8

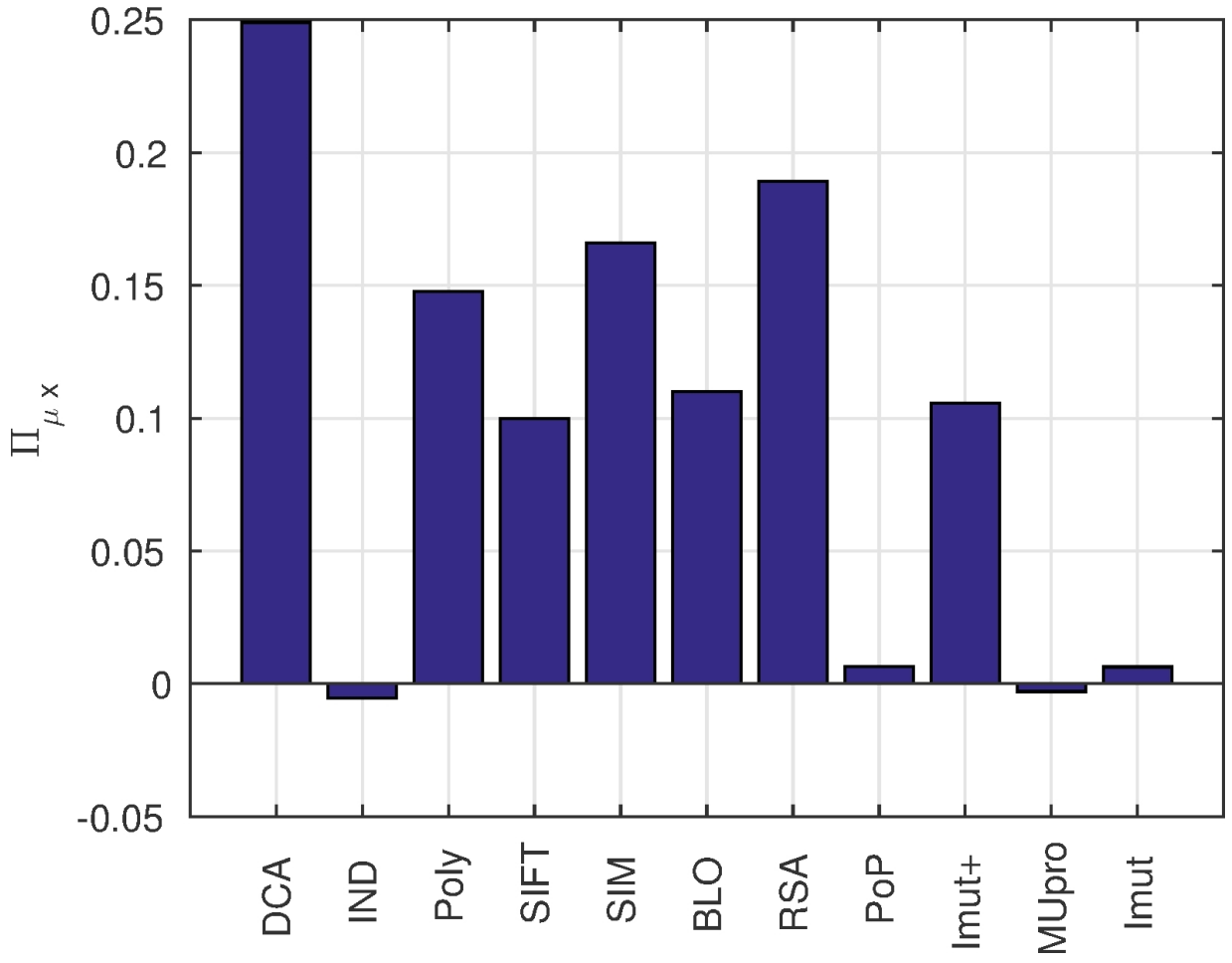

FIG. S8: **Partial correlations analysis of predictions for the beta-lactamase TEM-1:** Partial correlations $\Pi_{\mu x} = \omega_{\mu x}/\sqrt{\omega_{\mu\mu}\omega_{xx}}$ between the experimental fitness μ and a prediction method x, given all the others predictions, where ω is the inverse of the cross-correlation matrix ρ shown in Fig. S4. We find that the partial correlation between DCA and experimental fitness given all the others ($\Pi_{\mu,DCA} \approx 0.25$) is bigger than the partial correlations of any other method: This confirms that DCA provides important information about experimental fitness measures, which is not contained in the other descriptors.

Supplementary Figure 9

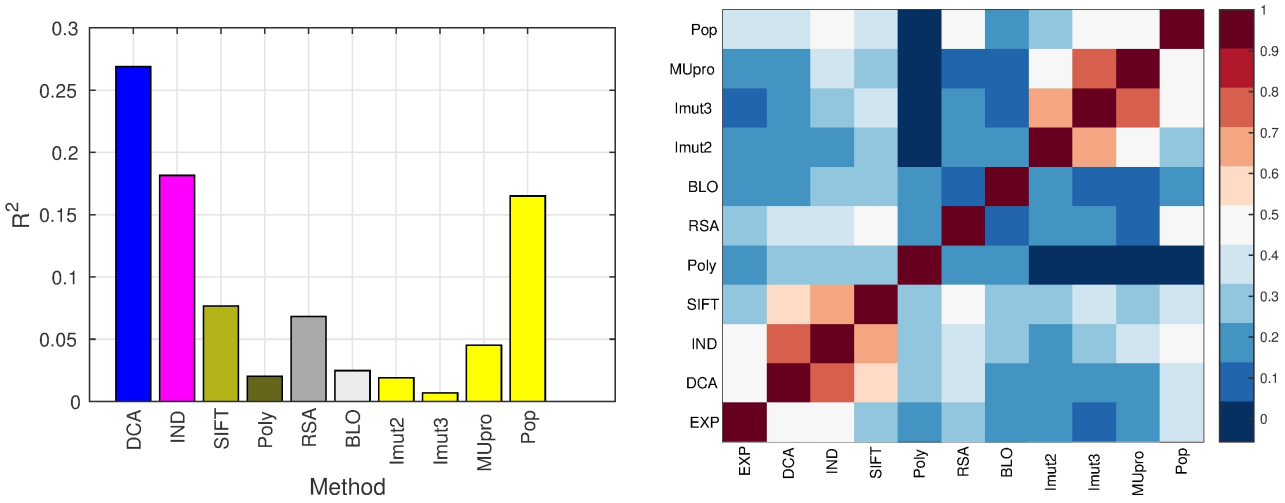

FIG. S9: **Analysis of the PDZ domain. Left Panel** – Correlation R^2 between experimental fitness (EXP) and computational methods for PDZ domain. Tested methods are DCA (DCA), Independent Statistical Model (Ind), SIFT (SIFT), Polyphen-2 (Poly), Relative Solvent Accessibility (RSA) , Blosum Substitution Matrix (BLO), PopMuSiC (POP), I-Mutant2.0 (Imut2), I-Mutant3.0 (Imut3) and MUpro (MUpro). **Right Panel** – Cross-correlations between methods.

Supplementary Figure 10

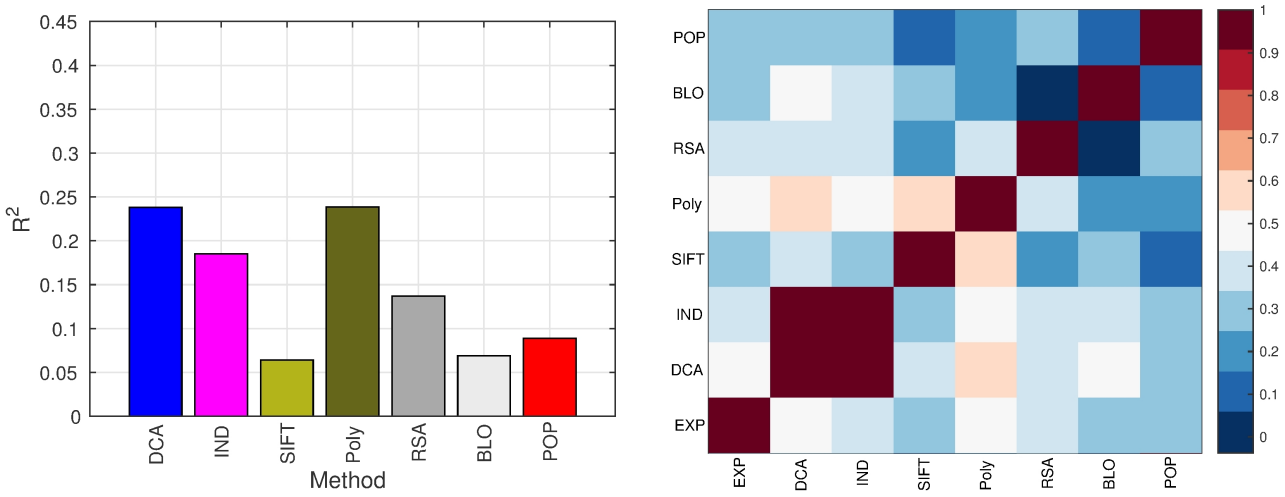

FIG. S10: **Analysis of the RRM domain. Left Panel** – Correlation R^2 between experimental fitness (EXP) and computational methods. Tested methods are DCA (DCA), Independent Statistical Model (Ind), SIFT (SIFT), Polyphen-2 (Poly), Relative Solvent Accessibility (RSA) , Blosum Substitution Matrix (BLO), PopMuSiC (POP). **Right Panel** – Cross-correlations between methods.

Supplementary Figure 11

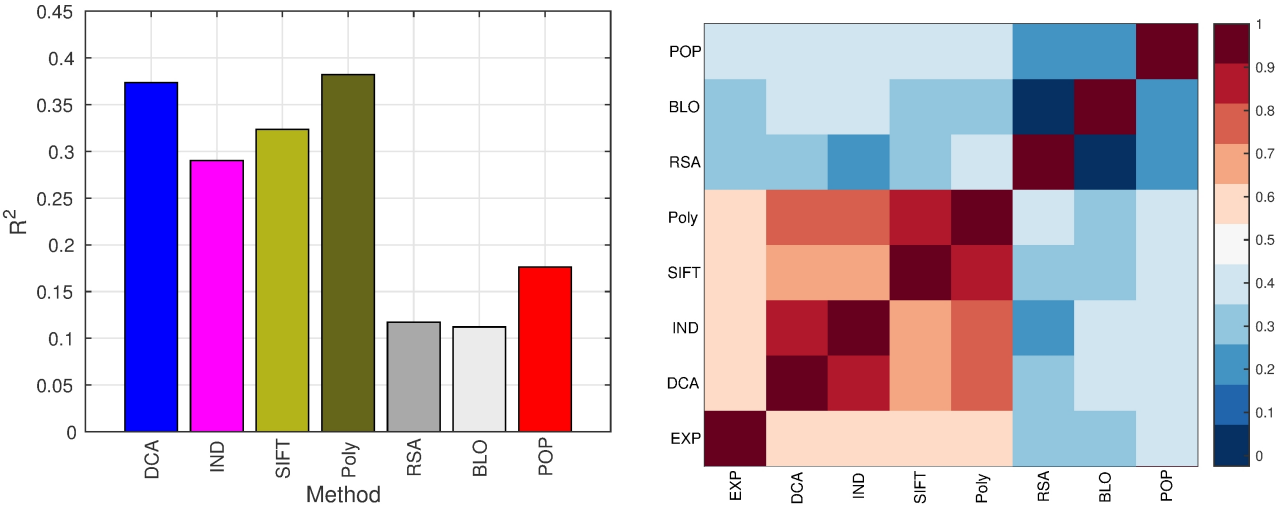

FIG. S11: **Analysis of the β-glucosidase enzyme. Left Panel** – Correlation R^2 between experimental fitness (EXP) and computational methods. Tested methods are DCA (DCA), Independent Statistical Model (Ind), SIFT (SIFT), Polyphen-2 (Poly), Solvent Accessibility (RSA) , Blosum Substitution Matrix (BLO), PopMuSiC (POP). **Right Panel** – Cross-correlations between methods.

Supplementary Figure 12

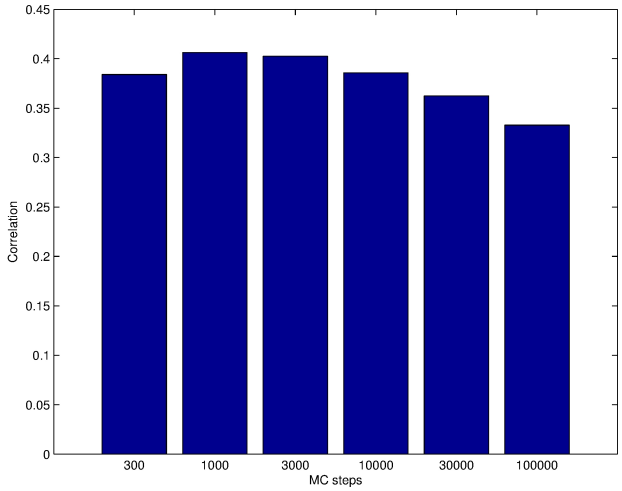

FIG. S12: **Correlation between experimental fitness and energetics for the beta-lactamase TEM-1** Pearson correlation between experimental fitness μ_{EXP} and $\mu(\phi_{stab})$, with $\Delta\Delta G$ computed at different number of MC steps.

Supplementary Figure 13

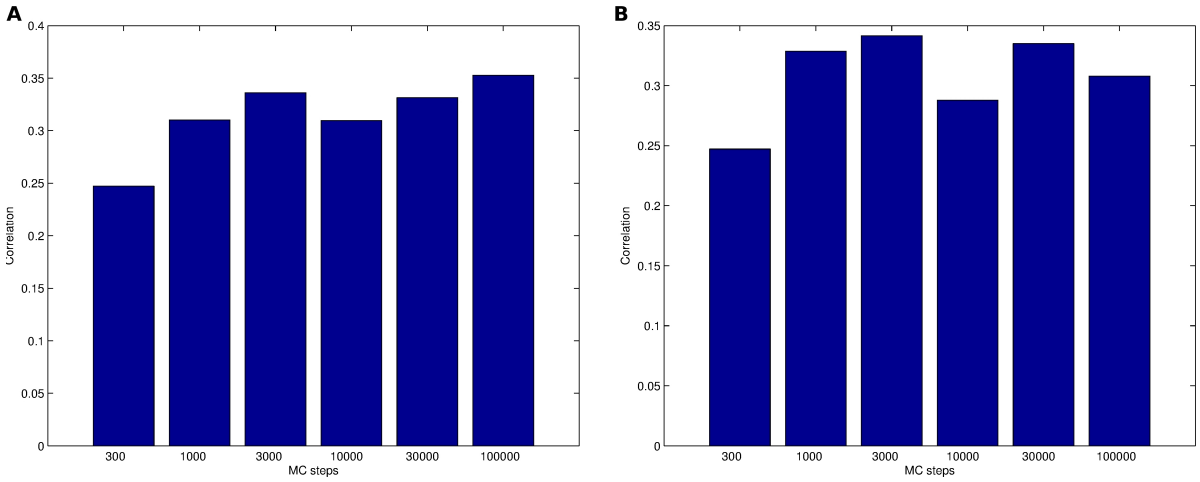

FIG. S13: **Correlation between experimental fitness and RMSD for the beta-lactamase TEM-1. Panel A** – Pearson correlation between experimental fitness μ_{EXP} and $\mu(\phi_{RMSD})$, with RMSD calculated on the whole protein and computed at different number of MC steps. **Panel B** – Pearson correlation between experimental fitness μ_{EXP} and $\mu(\phi^{EAS}_{RMSD})$, with RMSD restricted to the extended active site and computed at different number of MC steps.

Supplementary Text S1 - Prediction of the impact of single site mutations in different systems

To show the generality of our approach, we have applied it to predict the impact of single point mutations in three further different systems: a PDZ domain, an RNA recognition domain and the glycosidase enzyme.

PDZ domain – The mutational landscape of a member of the PDZ family domain ($PSD95^{pdz3}$), a common structural domain found in signalling proteins of a wide spectrum of organisms, has been recently characterized in [4] by quantitatively linking the ability of the protein to bind its cognate ligand (derived from CRIPT, a cysteine rich interactor of PDZ) to the expression of an enhanced Green Fluorescence Protein (eGFP). Among a bacterial population carrying large number of mutations in the protein under study, cells displaying eGFP levels above threshold have been selected and subsequently sequenced to assess the frequency of each allele in the selected and unselected populations.

RNA recognition domain – The effect of mutations on the RNA recognition domain RRM2 of *Saccharomyces cerevisiae* Pab1 protein (an essential poly(A)-binding protein) where studied by [5]: they deleted the wild-type gene and transfected two plasmids: a first one under the control of a tetracycline-off promoter containing the original gene and a second one constitutively expressed containing a variant of the gene. Yeast were grown to logarithmic phase and then diluted into tetracycline-containing media to shut off the expression of the wild type gene. The change in frequency was measured before (unselected) and after (selected) 22h of growth in the selection media.

Glycosidase enzyme – Using droplet microfluidic screening Romero et al. [6] have been able to map the activity of millions of sequence variants of Bgl3, a β-glycosidase enzyme from *Streptomyces*: a library of enzyme variants was expressed in *E.coli*, and single cell were encapsulated in microfluidic droplets containing lysis reagents and a fluorogenic substrate of the enzymes. Upon lysis, enzymes were released in the droplet and interacted with the substrate, and droplets containing efficient enzyme variants were sorted according to their brightness and then processed using next-generation sequencing.

Quantitative measure of mutation effect – The effect of each single point mutations has been quantified as the log frequency of observing each amino acid b at each position i in the selected (sel) versus the unselected (unsel) population, relative to amino acid a_i present in the wild type:

$$\mu_{exp}(a_i \to b) = \log\left[\frac{f_i^{sel}(b)}{f_i^{sel}(a_i)}\right] - \log\left[\frac{f_i^{unsel}(b)}{f_i^{unsel}(a_i)}\right] \tag{1}$$

Details of the model and analysis – Statistical inference and comparison with experimental fitness was performed identically as described in the Methods for the TEM-1 case, and we found that the explicative power of DCA is systemically higher than that of a non-epistatic independent model, as reported in table S4 together with other details of the analysis.

As done in the case of the inference of TEM-1 landscape, we have used DCA on reduced MSA, containing the residue position carrying the mutation of interest, and all residues, which are, in a representative TEM-1 crystal structure, within a distance d_{max}. Again we observe a rapid increase in predictive power when a structural neighborhood is taken into account, and the maximum correlation is reached around $d_{max} = 15 \sim 20\mathring{A}$ (data not shown).

Our statistical score outperforms other commonly used prediction tools (Fig. S9, S10 and S11): SIFT, Polyphen-2 and bioinformatic predictors for protein stability. We didn't perform molecular simulations in this case given their intensive computational cost.

System	PFAM	L	M	N	R^2_{DCA}	R^2_{IND}
PDZ domain	PF00595	81	3217	1426	0.27	0.18
RRM domain	PF00076	69	51991	1107	0.24	0.18
β-Glycosidase	PF00232	453	10563	2627	0.37	0.29

TABLE S4: **Details of the analysis**: The table shows the name of the PFAM family used in the analysis, the length of the multiple sequence alignment L, the number of sequences in the alignment M, the number of mutants in the test set N, and accuracy of predictions by DCA and by an independent model.

Supplementary refereces

[1] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1): 012707.

[2] Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution*, 31(6): 1581–1592.

[3] Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., *et al.* 2013. Capturing the mutational landscape of the beta-lactamase tem-1. *Proceedings of the National Academy of Sciences*, 110(32): 13067–13072.

[4] McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. 2012. The spatial architecture of protein function and adaptation. *Nature*, 491(7422): 138–142.

[5] Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. 2013. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly (a)-binding protein. *RNA*, 19(11): 1537–1551.

[6] Romero, P. A., Tran, T. M., and Abate, A. R. 2015. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences*, page 201422285.